Statistics Interpreted

R. D. Pavlovich
Graduate Instructor, Civil Engineering
Purdue University

INTRODUCTION

Most, if not all, highway designers and planners will, at one time or another, come in contact with statistics. Statistics is a potent tool that is rather well developed and the value of statistical methods is well proven; in other words, statistics is here to stay. All of us will use statistics directly or will use criteria based upon statistically obtained data and most probably have already encountered statistics in the current literature.

It is the intent of this paper to show some of the rudiments, discuss some of the more commonly used terms, and to provide a bibliography that can be of help if self study is desired. This short paper will not permit complete mathematical development of most of the concepts but the bibliography will provide any degree of rigor that most will require or desire.

To begin with, let us adopt some definition of statistics that is both useful and printable while recognizing that there are about as many definitions as there are authors. For our purposes let us define statistics as the science of making reliable generalizations based on sample data. In most cases rigorous mathematics is used to aid in assessing the validity of these generalizations. Statistics, as used in civil engineering planning and design, generally deals with the following problems among others:

1. Collecting and summarizing data and measurements.
2. Designing experiments and surveys.
3. Estimating population parameters (properties) from samples.
4. Testing hypotheses about population parameters (properties).
5. Studying relationships between variables (regression analysis and analysis of variance).

Finally, let us bear in mind that statistics operates on numbers only and not on physical phenomenon. The engineer must run the tool and not vice-versa. Final judgments and decisions are made by people; statistics simply aids in the decision process. We are all well aware that while figures don't lie . . . etc., etc. Consider for a moment the ad-

vertiser's and politician's standby, "nine out of ten prefer . . .," and ask yourself, before the final decision is made, "which ten did he observe and measure?" Lastly, along these lines, remember, that in the event one is working with a professional statistician in the design of experiments or in the analysis of data, to be certain that he fully understands the physical problem and is included in on the "whole thing" thus mitigating the problem of attempting to draw vast conclusions from half vast data.

Now, let us get on with some of the concepts of statistics and what they can and perhaps can not do for the planner and designer.

SOME BASIC DESCRIPTORS OF CENTRAL TENDENCY

Suppose that one wished to describe a group or groups of data. Say the data are as follows:

Group I	*Group II*
60	38
29	36
28	36
27	34

The idea is to provide a single number that will represent a group of individual values. There are several "averages" that can be used to describe these data. The most common of these averages is the *arithmetic mean* which is defined as:

$$\bar{x} = \frac{\Sigma x_i}{n}$$

where n is the number of individuals involved and x_i is an individual value.

The arithmetic mean for each group of data is:

$$\bar{x}_I = \frac{144}{4} = 36$$

$$\bar{x}_{II} = \frac{144}{4} = 36$$

Some quick reflection on these averages and on the data will show that while some information is compactly conveyed by the "mean," not much use can be made of it as yet. In fact, one could be in some trouble if the number 36 were his only guide as to the make-up of each of the above populations. For instance, which group of beauties would you select as being nearer a "perfect 36," if your only information was the mean and the number of individuals in each group?

Two more averages will be demonstrated that, in some cases, can be more descriptive than the arithmetic mean. These are the *median* and the *mode*. The median is the middlemost item or value whereas the mode is the most commonly occurring value.

For group I:

$$\bar{x} = 36$$

Median = 28½

Mode = (does not exist).

For group II:

$$\bar{x} = 36$$

Median = 36

Mode = 36

Two less used averages but occasionally useful are the *harmonic* and *geometric* means. The harmonic mean is defined as:

$$H = \frac{n}{\Sigma \frac{1}{x_1}}.$$

and is associated with rates. A good example (as given by Moroney) is as follows: consider a square of 100 miles per side and flight speeds of 100, 200, 300 and 400 miles per hour per side—see Figure 1. The average speed is *not:*

$$\frac{100 + 200 + 300 + 400}{4} \text{ or } 250 \text{ mph.}$$

It is however:

$$\frac{400 \text{ mi.}}{(1 + 1/2 + 1/3 + 1/4) \text{ hr.}} = \frac{400}{25/12} = 192 \text{ mph.}$$

Using the harmonic mean:

$$H = \frac{4}{\frac{1}{100} + \frac{1}{200} + \frac{1}{300} + \frac{1}{400}} = 192 \text{ mph.}$$

The geometric mean is useful when dealing with exponential situations such as population growth or compound interest types of problems. The geometric mean is defined as:

$$G = \sqrt[n]{x_1 \cdot x_2 \cdot \ldots \cdot x_n}$$

(See Figure 2)

In closing this section, it should be noted that there are at least five "averages," each useful in certain circumstances and in all cases it helps to know which one is being used or should be used.

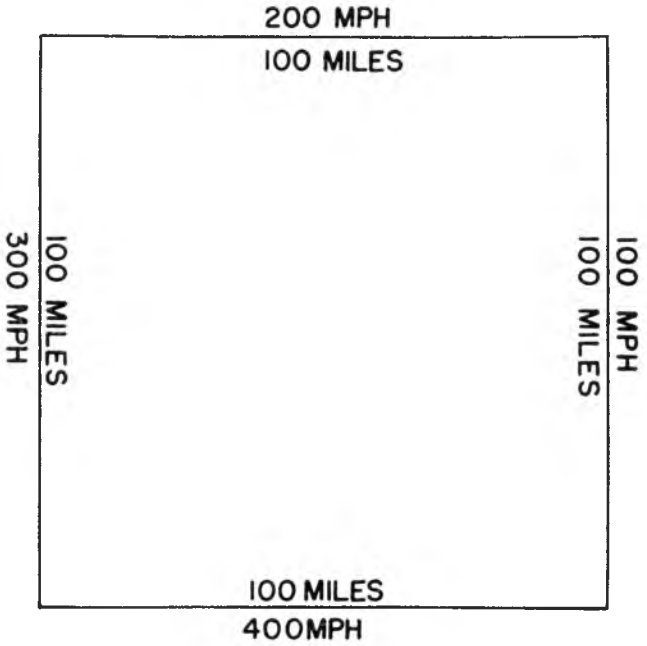

Figure 1

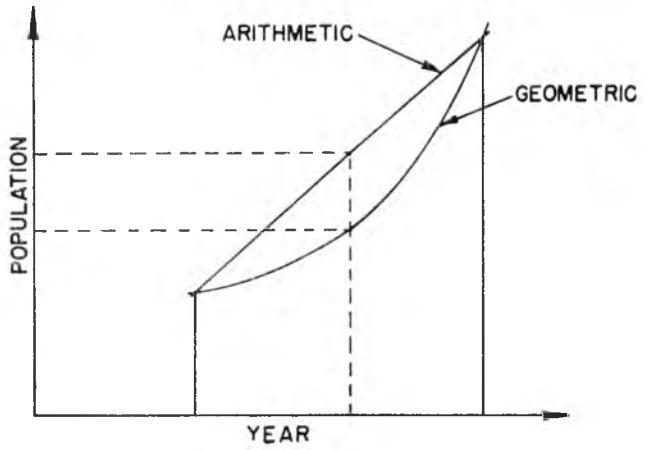

Figure 2

SOME CONCEPTS OF VARIABILITY

For our purposes we will define variability as the dispersion or scatter of values in the group under consideration.

If we recall the sets of data used before and that each has the same arithmetic mean, it is obvious that more information is desirable in order to more clearly decide which group is nearer the perfect 36. This additional information can come in the form of descriptors of variability.

Range (R) is defined as the highest minus the lowest value of the group. Advantage of the range is that it is quickly calculated (usually by inspection) but has the disadvantage that it tends to magnify the extreme values. For the data groups given:

$$R_I = 60—27 = 33$$
$$R_{II} = 38—34 = 4$$

This tells us that while both groups have the same mean (36) that group II has less scatter than group I (which is patently obvious when examining the data but may not be so obvious if there were larger amounts of data in each set or group). Observe the increase of information about the population by including, with the mean, a single number that defines, however inadequately, variability of the group of data.

Another measure of variability that is somewhat more useful than the range is the *standard deviation* of the sample or population and is defined as the root-mean-square of the deviations from the mean. The formulation for standard deviation is as follows:

$$s = \sqrt{\frac{(x_i—\bar{x})^2}{n—1}}$$

$$\sigma = \sqrt{\frac{(x_i—\bar{x})^2}{n}}$$

where s is the standard deviation of a sample and σ is the standard deviation of a population. Note that the difference between the sample and the population is n vs (n—1) in the denominator. Space does not allow a discussion of the reasons for this difference, however most of the references will give the matter full and adequate discussion. Note that the standard deviation is almost an average deviation from the mean for a set of data.

A very important measure of variability that bears great significance in mathematical development of statistical formulations is the *variance* and is defined as the square of the standard deviation. The formulation for the variance is:

$$\sigma^2 = \frac{\Sigma (x_i - \bar{x})^2}{n} \quad \text{for a population}$$

$$\text{or } s^2 = \frac{\Sigma (x_i - \bar{x})^2}{n-1} \quad \text{for a sample,}$$

Some algebraic manipulation provides a little more workable equation for machine calculation and is given by:

$$s^2 = \frac{\Sigma x_i^2 - \frac{(\Sigma x_i)^2}{n}}{n-1} = \frac{n \Sigma x_i^2 - (\Sigma x_i)^2}{n(n-1)}$$

When making these calculations please observe that Σx_i^2 means to square each value then take the sum of all these squares whereas $(\Sigma_i)^2$ means to obtain the sum of the values then square this sum.

For each of the data groups used in the previous example, the following tabulations are provided:

Group I

x_i	$(x_i - \bar{x})$	$(x_i - \bar{x})^2$
60	24	576
29	−7	49
28	−8	64
27	−9	81
Σ 144	0	770

$R = 60 - 27 = 33$

$\sigma^2 = \Sigma(x_i - \bar{x})^2/n = 770/4 = 192.5$

$\sigma = 192.5 = 13.86$

Group II

x_i	$(x_i - \bar{x})$	$(x_i - \bar{x})^2$
38	2	4
36	0	0
36	0	0
34	−2	4
Σ 144	0	8

$R = 38 - 34 = 4$

$\sigma^2 = \Sigma(x_i - \bar{x})^2/n = 8/4 = 2$

$\sigma = 2 = 1.414$

DISTRIBUTIONS

Without thorough discussion let us define a distribution as a function that describes the frequency of occurrence of a particular event or value. There are three distributions commonly encountered in current literature; the hypergeometric, binomial and the normal distributions. Only a short description will be given of the hypergeometric and binomial situations.

Consider first, the *combination* of r elements from a set of n possibilities.

$$C_r^n = \frac{n!}{(n-r)!\ r!}$$

As an example, how many ways are there to select five items from a group of ten?

$$C_5^{10} = \frac{10!}{(10—5)!\ 5!} = \frac{10!}{5!\,5!} = 252$$

Now, the *hypergeometric* distribution results from sampling *without* replacement from a population that contains n_1 successes and n_2 failures (Note that $n_1 + n_2 =$ entire population). This distribution is described by:

$$P(x) = \frac{C_x^A\ C_{n-x}^B}{C_n^{A+B}}. \quad \text{where:}$$

$P(x)$ = probability of success
x = number of successes in a sample of size n.
A = number of successes in the population from which the sample was taken.

B = number of failures

Consider the following: What is the probability of three aces in the first five cards dealt?

$x = 3$
$A = 4$
$B = 48$
$n = 5$

$$P(x) = \frac{C_3^4\ C_2^{48}}{C_5^{52}}. = 1.732(10)^{-3}$$

which shows chances are fairly poor that the first three players in a five man game will have an ace in the hole. But then, perhaps, practice makes perfect.

Whereas the hypergeometric distribution resulted from sampling without replacement the *binomial* distribution involves sampling *with* replacement. This distribution is given by:

$P(x) = C_x^n \; P^x \; (1-p)^{n-x}$ where:

$x =$ the number of successes in the sample size n.

$P =$ probability of a success.

It will remain as the well known exercise for the student to use the example given in the hypergeometric case and to apply the binomial probability equation remembering to sample with replacement.

The most commonly used distribution function is the well known *normal* distribution. The properties are well known and tabulated in all the references. Most natural occurrences are distributed normally or can be simply transformed to fit this function and hence, most test statistics are based on this curve.

This normal curve is given by:

$$Y = \sigma \sqrt{\frac{1}{2\pi}} \cdot \exp - \left[\frac{(x-\mu)^2}{2\sigma^2}\right]$$

where x is the value or event under consideration and Y is the frequency of occurrence of the event μ is the arithmetic mean of the population and σ^2 is the population variance. Note that μ and σ^2 are properties or parameters of the function.

If the variable x is replaced by:

$$z = \left(\frac{x-\mu}{\sigma}\right)$$

and if $\mu = 0$, $\sigma = 1$ the "standard form" of the normal distribution becomes:

$$y = \frac{1}{\sqrt{2\pi}} \quad \exp - \left[\frac{z^2}{2}\right]$$

(See Figure 3)

This normal curve is a probability curve and in the standard form the area under the curve is 1.0. Hence the probability of x falling between $-\infty$ and $+\infty$ is 1.0. One final observation that is useful when dealing with means and standard deviations; 68 percent of the

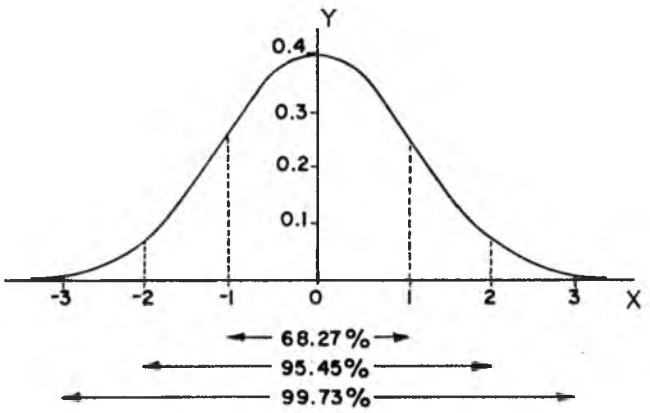

Figure 3

area lies between the mean and ± 1 standard deviation and 99.8 percent of the area lies between the mean and ± 3 standard deviations. Hence, for instance, if examination scores were normally distributed and if you were given the class average and variance (or standard deviation) you would be able to determine what percentage of your colleagues or competitors who were above or below your score.

SAMPLING

As was stated earlier, one of the uses of statistics is to infer the properties of a population from a relatively small sample. We would like, for instance, to infer the strength of concrete in a structure from a relatively small sample. We must sample because it is impossible to test all of the concrete.

To begin with: If a sample is biased and not "representative" of the population any conclusions made regarding that population will most likely be erroneous. This is just common sense and simple and is probably why the concept is forgotten (sometimes by convenience) or ignored. Remember the grain of salt involved in the case of, "Nine out of ten prefer . . ." There must be some random sampling technique employed that will insure that each value in the population has an equal chance of being chosen and placed in the sample; purely, and simply, period.

A common symbol convention, and one that is now almost standard in most of the literature is to use Greek symbols for population parameters and Roman letters for sample parameters.

	Population	*Sample*
Mean	μ	$\bar{x}$
Standard Deviation	σ	s
Variance	σ^2	s^2

Now, to demonstrate the type of information a sample can give us about a population and how one may interpret this information, let us again consider data group I (60, 29, 28, 27, $\mu = 36$, $\sigma^2 = 192.5$). Note that these data are not particularly well behaved nor "nice" and normal. Let us consider sampling all possible combinations of two (sampling with replacement). The following table cells show the mean of the two values on top and the standard deviation below (note that s is calculated on the basis of n − 1 or 1).

	60	29	28	27
60	60 0	44.5 21.92	44 22.63	43.5 23.33
29	44.5 21.92	29 0	28.5 0.71	28 1.41
28	44 22.63	28.5 0.71	28 0	27.5 0.71
27	43.5 23.33	28 0.71	27.5 0.71	27 0

Consider now, the means:

$$\Sigma x_i = 576.0$$
$$\Sigma x_i^2 = 22{,}276$$
$$(\Sigma x_i)^2 = 331{,}776$$

The mean of the means is $\bar{x}_{\bar{x}} = \frac{576}{16} . = 36.0$; things should begin to take shape now because $\bar{x}_{\bar{x}} = \mu = 36.0$. Now calculating the variance of the means and we find $\sigma_{\bar{x}}^2 = 96.25$

If we note the following:

$$\frac{\sigma^2}{\sigma_{\bar{x}}^2} = \frac{192.5}{96.25} = 2.000$$

We then see that $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$.

where $\sigma_{\bar{x}}$ is called "standard error of the mean".

Calculation of the mean of the standard deviations gives 140.72/16 = 8.795 which does not provide much information regarding properties of the population but if the mean of the variances is calculated as 3078.5/16 = 192.406 it is seen that this value compares quite well with the variance of the population.

INFERENCES

It is apparent from the last section that a sample variance and sample mean can be used to estimate, with some degree of confidence, these population parameters. It should also be apparent that the precision of the estimate is a function of the number or size of the sample. Based on the properties of the normal curve then, one can predict the limits of the parameters.

There are three statistics commonly used in the current literature to set limits on the predicted value of a parameter; "Student's t test for means," "F test for variance," and "x^2 tests for goodness of fit" that will be shown without a good deal of discussion:

$$t = \frac{\bar{x} - \mu}{s/\sqrt{n}}$$

"t" has a distribution something like the normal distribution with properties tabulated such that the area under the curve provides a probability function—see Figure 4.

Let us say that we have a sample mean x, and sample standard deviation s, and wish to know if this sample comes from a population with

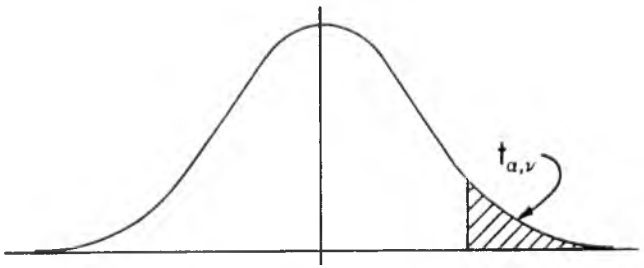

Figure 4

mean μ. Let us also say that we wish to be 95 percent confident (α) of our answer, i.e., 95 percent of the time we are correct. After calculating t we compare our value with a tabulation based on α and the degrees of freedom ($\upsilon = n - 1$) and note that there is a 5 percent probability that our calculated t should exceed the tabulated value. If our calculated value is less than the tabulated we will conclude that the sample is from the population with mean μ. We could place any confidence limit α we wish but as we become more and more confident we require a larger sample. As an example let us attempt to set limits on the mean of the population for the data of group I based on, say, observations 28 and 27. Some manipulation of the equation for t gives:

$$\mu = \bar{x} \pm t \frac{s}{n}$$

For 95 percent confidence, $t_{0.95,\ \upsilon = 1} = 12.706$

$\bar{x} = 27.5$ $\qquad$ $s = 0.707$

$s^2 = 0.5$ $\qquad$ $n = 2$

$$\mu = 27.5 \pm (12.706)\frac{0.707}{2}$$

$$= 27.5 \pm 6.4$$

Hence: $\qquad 21.1 < \mu < 33.9$

But, you say, we *know* that the population mean is 36 but statistics shows that it is somewhere between 21.1 and 33.9. We also know we will be wrong 5 percent of the time; statistics has at least placed limits on our uncertainty. If, however, we had chosen a 99 percent confidence level ($t = 63.66$) we would predict μ as being between -4.3 and 36.2. In this case we are more certain but the "band of prediction" has widened considerably.

The F test and the x^2 goodness of fit test work in about the same way; calculate the statistic and compare with a tabulated value.

F tests the typothesis that two sample variances are from the same population with

$$F = \frac{s_1^2}{s_2^2}$$

and the goodness of fit test is based on

$$x^2 = \Sigma \frac{(O—E)^2}{E} . \qquad \text{where:}$$

O = observed data

E = the expected value (based on the normal or whatever distribution is in question).

LEAST SQUARES FIT FOR LINEAR DATA (REGRESSION)

The last topic before closing is that of fitting a straight line through a set of data that exhibits some scatter.

Consider the straight line:

$$Y = b + mx$$

or $\hat{Y} = \beta_o + \beta_1 X_i$ (See Figure 5)

Now add some data points (See Figure 6)

where: X_i = observed X value

Y_i = observed Y value

$\hat{Y}$ = predicted Y value from $\hat{Y} = \beta_o + \beta_1 X_i$

The "best fit" line through this data is the one that has minimum error or residuals.

$$Y_i = \beta_o + \beta_1 X_i + \epsilon_i$$
$$\epsilon_i = (Y_i - \beta_o - \beta_1 X_i)$$

"Sum of squares" of error: (SSE)

$$S = \Sigma\epsilon_i^2 = \Sigma(Y_i - \beta_o - \beta_1 X_i)^2$$

To minimize SSE:

$$\frac{\partial S}{\partial \beta_o} = -2\Sigma (Y_i - \beta_o - \beta_1 X_i) = 0$$

$$\frac{\partial S}{\partial \beta_1} = -2\Sigma X_i(Y_i - \beta_o - \beta_1 X_i) = 0$$

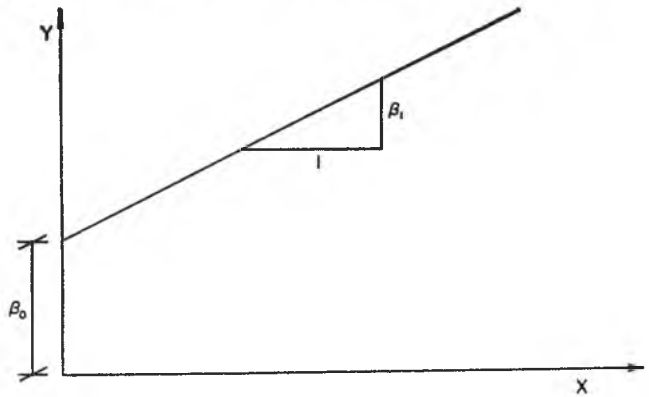

Figure 5

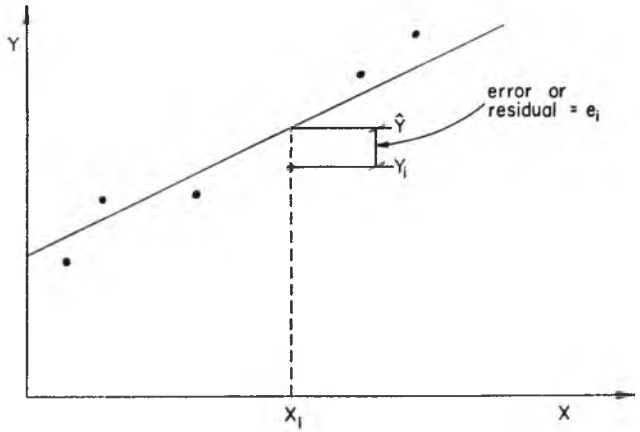

Figure 6

"Estimate" (β_o, β_1) by b_o, b_1), then:

$$b_o n + b_1 \Sigma X_i = \Sigma Y_i \qquad (I)$$

$$b_o \Sigma X_i + b_1 \Sigma X_i^2 = X_i Y_i \qquad (II)$$

Solution of (I) and (II) gives:

$$b_1 = \frac{\Sigma X_i Y_i - [(\Sigma X_i)(\Sigma Y_i)]/n}{\Sigma X_i^2 - \frac{(\Sigma X_i)^2}{n}} = \frac{\Sigma (X_i - \overline{X})(Y_i - \overline{Y})}{\Sigma (X_i - \overline{X})^2}$$

$$\& \; b_o = \overline{Y} - b_1 \overline{X}$$

where $\overline{Y}$ and $\overline{X}$ are the means of the Y and X sets of data. Hence, a simple calculation of b_1 and b_o provides the best fit.

Two methods of determining the quality of fit are commonly used; the "correlation coefficient (r) and the "correlation ratio" (R^2) should both be very close to unity for a perfect fit.

The correlation coefficient is given by

$$r = \frac{1}{n} \Sigma \left(\frac{X_i - \overline{X}}{\sigma_x}\right)\left(\frac{Y_i - \overline{Y}}{\sigma_y}\right)$$

where:
the notation is the same as has been previously discussed and provides results along these lines—see Figure 7.

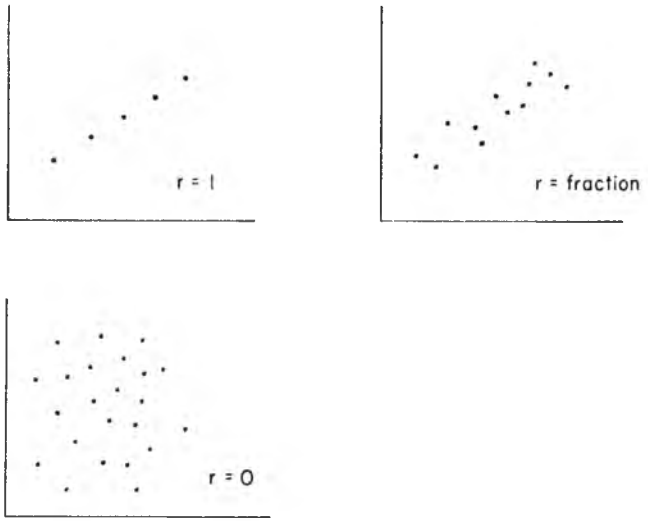

Figure 7

The concept of the correlation ratio is as follows—see Figure 8. Using the sum-of-squares technique:

$$\begin{array}{lccccc} & \text{SS} & = & \text{SS} & + & \text{SS} \\ & \text{about} & & \text{about} & & \text{due to} \\ & \text{mean} & & \text{regr.} & & \text{regr.} \end{array}$$

$$\text{or} \quad \Sigma(Y_i - \overline{Y})^2 = \Sigma(Y_i - \hat{Y})^2 + \Sigma(\hat{Y} - \overline{Y})^2$$

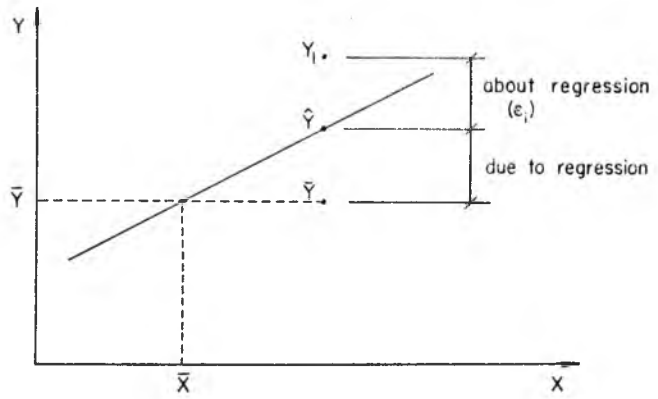

Figure 8

To minimize the error it would be good if ϵ_1 or SS about regression would be as near zero as possible. Then: $\Sigma(Y_i-\overline{Y})^2 = \Sigma(\hat{Y}-\overline{Y})^2$

$$\text{or } R^2 = \frac{\Sigma(\hat{Y} - \overline{Y})^2}{\Sigma(Y_1 - \overline{Y})} = 1$$

Conclusion

It is hoped that this discussion, necessarily brief due to time and space requirements, has shown where some of statistical terminology encountered in the literature comes from and what it means. More importantly, it was intended to show that the subject follows some degree of logic and that it is not beyond the comprehension and use of present designers and planners. By using the selected references and investing some time in study and mastering only a few concepts and methods an extremely potent tool can be included in the engineer's techniques to more efficiently solve present and future planning and design problems.

REFERENCES

1. Dixon, W. J. and Massey, F. J., "Introduction to Statistical Analysis," McGraw-Hill. Good treatment of elementary statistics, easy to read and follow and has worked examples.
2. Miller, I. and Freund, J. E., "Probability and Statistics for Engineers," Prentice-Hall, Inc. This is the present text for a first course at the graduate level at Purdue.
3. Downie, N. M. and Heath, R. W., "Basic Statistical Methods," Harper and Row. Intended for the beginning student of social sciences. Not as "mathematical" as reference 2.
4. Ostle, Bernard, "Statistics in Research," The Iowa State University Press. An excellent book for those with some background. Contains procedures and formulas, etc. for some of the more involved analysis techniques.
5. Draper, N. R. and Smith, H., "Applied Regression Analysis," John Wiley and Sons, Inc. An excellent development of regression concepts by Matrix Methods (which, by the way, are not difficult and greatly simplify matters). Introduces solutions of involved problems by computer methods and discusses several computer techniques of modeling for several variables.
6. Mandel, J., "The Statistical Analysis of Experimental Data," Interscience Publishers. A very thorough treatment of the ideas and concepts underlying modern theory; an extremely clearly and

well written book that makes you appreciate a good practical mathematician.

7. Crow, E. J., Davis, F. A. and Maxfield, N. W., "Statistical Manual," Dover Publications (paperback). Originally the Naval Ordnance Manual. A good "cookbook" with methods and formulas.
8. Spiegel, M. R., "Statistics," Schaum's Outline Series, McGraw-Hill. A typical Schaum's outline that contains a wealth of problems with solutions and/or answers. Highly recommended for self-study.
9. McLaughlin, J. F. and Hanna, S. J., "Evaluation of Data," Reprinted from American Society for Testing and Materials STP 169-A (Significance of Tests and Properties of Concrete and Concrete Making Materials). This is "Purdue University Engineering Reprints No. CE 222", and is available at the Road School reprint room.
10. Moroney, M. J., "Facts From Figures," Pelican paperback. A good and fairly light treatment of general statistics with considerable English common sense philosophy included. Highly recommended.